Published in: Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics (QITl-6. Tübingen: Universität. 2015.

Reading corpora as an instrument for studying a relevance-based account of language processing

A case study using a reading corpus of German jurisdictional texts

Sascha Wolfer
Institut für Deutsche Sprache
Mannheim, Germany
wolfer@ids-mannheim.de

***Abstract*—Reading corpora are text collections that are enriched with processing data. From a corpus linguist's perspective, they can be seen as an extension of classical linguistic corpora with human language processing behavior. From a psycholinguist's perspective, reading corpora allow to test psycholinguistic hypotheses on subsets of language and language processing as it is 'in the wild' – in contrast to strictly controlled language material in isolated sentences, as used in most psycholinguistic experiments. In this paper, we will investigate a relevance-based account of language processing which states that linguistic structures, that are embedded deeper syntactically, are read faster because readers allocate less attention to these structures.**

Keywords—language processing; reading; corpora

I. INTRODUCTION

In linguistics, many research efforts are driven by or based on large collections of texts. The area of corpus linguistics revolves around these text collections and how they can be used to answer questions about language. One important aim when compiling corpora is to include data that is ecologically valid, i.e. language as it is produced and received by a large number of language users within the area of interest. In brief, corpus linguists want to include (subsets of) language as it is 'in the wild' into our corpora.

In psycholinguistics, research is mostly based on language material that is more distant to language as it is used 'in the wild'. There are good reasons for that. In most psycholinguistic studies, 'hand-crafted' stimuli sentences are used that are tailored to the specific research question at hand. If we are interested, for example, in attachment preferences of prepositional phrases (PPs), we may only alter the PP (or any other element which we suspect to have an influence on PP attachment) in a sentence while the rest remains constant. Other stimuli sentences then are a variety of the same sentence construction. In this way, psycholinguists hope to minimize noise and avoid confounds with other linguistic variables like word frequency or familiarity, abstractness, and many more. The progress that has been made in the field of psycholinguistics over the last decades proves that this approach cannot be wrong.

In the last decade, a hybrid approach between classic linguistic corpora as collections of text and traditional psycholinguistic stimulus material was developed. So-called eye-tracking corpora or reading corpora combine the collection of language as it was produced in the wild with psycholinguistic research methods. From the corpus linguistic side, this idea can be described as adding another annotation level to an existing text collection. This annotation level contains information about the processing times of words or other linguistic elements like phrases or sentences from many different participants. For each word in the corpus, we can then look up how long each participant's eye rested on a specific word. If we cross-combine this information with other annotations (e.g., the kind of phrase(s) a word is located in), we can abstract from single words and single participants. This is absolutely necessary because we do not want to draw conclusions like 'most participants read word *a* longer than word *b*'. What we want to conclude is something like 'on average and after taking word length and word frequency into account, verbs are read longer than nouns'.

From the perspective of a psycholinguist, a reading corpus may be described as a collection of processing data (mostly eye-tracking data, but also see Frank et al., 2013) on language material that is not explicitly constructed for the use in one specific experiment that is done to answer one specific research question. We have to bear in mind, though, that there are many research questions that may never be answered only with reading corpora. Mostly because there are interesting linguistic phenomena that appear so seldom in language as it is used in the wild that it is very unlikely that a relevant construction is indeed included in a reading corpus that was constructed for a more general set of research questions. One example of such a construction is a phenomenon called local syntactic coherences (cf. Tabor, Galantucci, & Richardson, 2004). This term is used to describe sub-structures within sentences that are themselves perfectly fine sentences. For example, in the sentence 'The coach chided the player tossed a frisbee by the opposing team' there is a main clause 'hidden' that says 'the player tossed a frisbee', which says exactly the opposite of the whole sentence. If we would want to investigate the processing of such structures in a reading corpus we might just run into the problem of data sparsity because natural language simply contains to few instances of local syntactic coherences. From a

corpus linguist's viewpoint, reading corpora also have some limitations that have to be pointed out. It might be very interesting to couple language comprehension data with corpora. However, we have to make compromises in terms of size. Compared to 'normal' corpora that do not contain any human processing information, it is very expensive to collect reading corpora because we have to invite people to the lab and collect data of them reading the texts. The larger the corpus gets, the more material a single person has to read. That is why the *Dundee Corpus* (Kennedy, Hill, & Pynte, 2003), the first reading corpus available, contains 'only' information for roughly 100,000 tokens (56,212 English tokens taken from *The Independent* and 52,173 French tokens from *Le Monde*). Each of these tokens has been read by 10 participants. For German, there are two reading corpora available that are quite different in their orientation. The *Potsdam Sentence Corpus* (Kliegl, 2004) contains 1,138 tokens read by 222 participants in isolated sentences that were constructed around specific target words. These target words have been selected for their word class, corpus frequency, and length. So, the Potsdam Sentence Corpus can be thought of being very close to the highly controlled psycholinguistic stimuli because its sentences do not bear any connections between them on the discourse level. The other German reading corpus is the PopSci Reading Corpus (Wolfer et al., 2013). It contains roughly 20,000 tokens from texts that were taken from German popular science journals. Here, sentences were presented in their original order, enabling the researchers to investigate phenomena on the discourse level. Of course, the use of naturalistic language material comes with the price of less control over the stimuli.

We are going to present a study using a specialized reading corpus of German. It is called the *Freiburg Legalese Reading Corpus* because data was collected at the University of Freiburg and it contains processing data for jurisdictional texts. Most of the text material was taken from decisions by the Federal Constitutional Court (Bundesverfassungsgericht) of Germany. In this paper, we will focus on a question regarding processing of naturalistic texts. We will present our research question and hypotheses in the next chapter. In chapter 3, we will present the corpus annotations we use and go into detail about the data selection. Chapter 4 will present the results of our analyses. In chapter 5, we will sum up and discuss the results before chapter 6 will present future research possibilities.

II. Questions & hypotheses

Generally speaking, content words that are embedded deeper in the syntactic structure tend to be read faster. Pynte, New, and Kennedy (2008a, 2008b) report this speed-up effect for deeper embedded content words in first-pass reading times – the duration between first entering a word *n* until the gaze is shifted to a word left or right to word *n*. They propose that this is the result of a reading strategy: "[D]eeply embedded words are frequently in a position of a modifier. They are more likely to function as members of a prepositional phrase (PP), an adjectival phrase, a relative clause, etc., and will, by definition, be less central to the main topic of the sentence than less deeply embedded words. For this reason, they may receive less attention, with less time devoted to semantic integration processes" (Pynte et al., 2008a, p. 8). The term 'strategy' suggests a conscious process of resource allocation by the reader. This does not have to be the case. Pynte et al. (2008a) do not state this explicitly, but such a strategy could well be learned by readers during their prior reading experience. This would mean that it is more of an automatic process that allocates attentional resources dependent on the supposed relevance of the currently read language material. One clue for low relevance would then be a deep syntactic embedding of a constituent. However, this does not hold for all kinds of constituents or words alike. The effect should be visible for modifying structures like adjectives and adverbs. Also, deeper embedded nouns should be read faster – especially if they are in modifying phrases (adjectival phrases, adverbial phrases or PPs).

So, to sum up, our overall research questions and connected hypotheses are: Are words in the Freiburg Legalese Reading Corpus, that are embedded deeper in the syntactic structure, really read faster? And if so, does this hold for all words? Relevance-based accounts suggest that this effect should be especially visible for words that are parts of modifying structures.

III. Annotation & data selection

The Freiburg Legalese Reading Corpus is annotated with phrase structure and part-of-speech information. Each word is a leaf in the phrase structure tree of its sentence. So, depth of embedding of a word is operationalized as the number of parent nodes in the phrase structure tree. We can use the part-of-speech information to determine different effects of depth of embedding on different (sets of) parts-of-speech. Our hypothesis suggests that adjectives and adverbs should be effected more strongly than, for example, finite verbs.

We selected all content words from the Freiburg Legalese Reading Corpus. These are 116,081 instances of read words. Associated information with each word is its part-of-speech, its depth of embedding and all reading time measures that are available in the corpus. In this contribution, we will only analyze first-pass reading times, already described above.

Further information that is associated with each word is its length in characters as well as its token frequency and orthographic familiarity extracted from dlexDB, a large corpus collection of German texts (Heister et al., 2011). Orthographic familiarity is operationalized as the cumulated frequency of all words with the same initial character trigram and the same length as word *n*, including word *n* itself.

IV. Analyses & Results

We used first-pass reading times as our criterion (or, in experimental terminology, our dependent variable). We used linear-mixed effects regression models from the R (R Core Team, 2015) package lme4 (Bates et al., 2015) to fit the regression models. Mixed models allow for the inclusion of random effects into the predictor structure of the model. In our case, we included a random intercept for participants because we wanted to control for the fact that some participants are generally reading slower or faster than others. Word length, token frequency, and orthographic familiarity were included as covariates to control for effects on the lexical level. We also included simple effects of the relative position of the word in

its sentence and its residual depth of embedding. Also, the interaction between the two predictors was included. Depth of embedding was residualized by the relative position of the word in the sentence to de-correlate the two variables. All variables were centered before including them into the model. Results of the model are summarized in Table 1. Whenever the absolute value of a *t*-value is above 2, we can be sure that the effect of the respective predictor is significant. The signs of the estimate and the t-value gives us the direction of the effect.

TABLE 1: MODEL RESULTS FOR ALL WORDS

Variable	Estimate	Std.Err.	*t*-value
Word length	0.041	0.0005	84.2
Token frequency	-0.056	0.0022	-25.2
Orthographic familiarity	0.041	0.0029	14.2
Relative position	-0.033	0.0061	-5.40
Depth of embedding	-0.014	0.0010	-13.6
Relative position X Depth of embedding	-0.007	0.0036	-1.99

As expected, all effects on the lexical level show significant effects on first-pass reading times. The more characters a word has, the longer it is read. The more frequent a word is in the language, the shorter it is read. The more familiar a word is, the longer it is read. This last effect could come as a surprise. However, it has to be seen in context with the effect of token frequency. Token frequency and orthographic familiarity are correlated, because the frequency measure of word *n* is part of the familiarity measure. If this effect is already included in the model, as in our case, then only the cumulated frequency of lexical competitors (words of the same length and the same initial character trigram as word *n*) is still available to explain variance in reading times. So, the positive effect of orthographic familiarity can be thought of as an effect of lexical competition.

The effects of the relative position of the word and its depth of embedding point into the expected directions. Words, that are embedded deeper, are read faster. The interaction, though quite close to the critical value of 2, suggests that the effect of depth of embedding is stronger for words that are closer to the end of a sentence. In one hypothesis, we stated that the effect of depth of embedding should be stronger for some parts-of-speech than for others. When splitting analyses for verbs, finite verbs, nouns, and adjectives/adverbs, the effect of embedding depth can be shown for each group. The effect seems to especially strong for nouns (due to limited space, these models are not supplied here).

So, to further investigate this effect, we concentrated on nouns. Since phrase structure annotation is available, we can annotate each noun with the information if it is directly embedded within a PP, i.e. if the node directly above the noun in the phrase structure is a PP. PPs can be thought of as a typical modifying structure. If modifying structures really are read faster with increasing depth of embedding, we would except an interaction effect. If a noun is located in a PP, the speed-up effect of syntactic embedding should be even more pronounced than for all other nouns. We did not only include nouns directly embedded in PPs, but also nouns that are directly embedded in coordinated prepositional phrases (CPP). In this analysis we are left with 53,043 nouns. 17,826 (33.6 %) of those have a PP or a CPP as their direct parent node in the phrase structure annotation.

TABLE 2: MODEL RESULTS FOR NOUNS

Variable	Estimate	Std. Error	t-value
Word length	0.044	0.0006	73.2
Token frequency	-0.062	0.0031	-20.0
Orthographic familiarity	0.021	0.0041	5.15
Relative position	-0.060	0.0091	-6.59
Depth of embedding	0.010	0.0019	-5.12
in (C)PP	-0.022	0.0060	-3.93
Depth of embedding X in (C)PP	-0.009	0.0030	-2.88

Table 2 shows that all effects previously shown for all words also apply for nouns: Long nouns are read longer, highly frequent nouns are read faster, familiar nouns are read longer, the relative position of the noun in its sentence has a negative effect on reading times and so does the depth of embedding. The last row shows that the embedding effect is modulated by the fact if a word is directly embedded in a PP. Figure 1 gives an impression of this interaction effect. It basically shows that the embedding effect is stronger for nouns that are located in PPs than for nouns that are located in other phrase types.

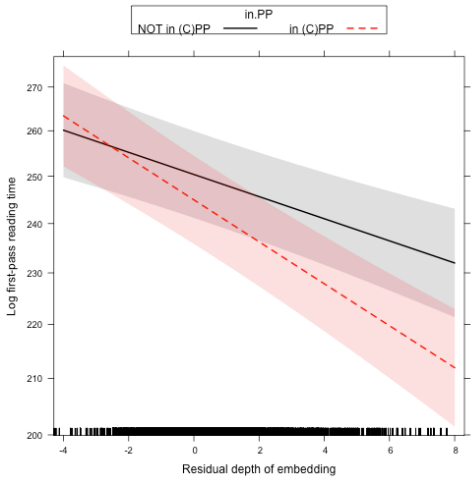

FIGURE 1: INTERACTION EFFECT BETWEEN DEPTH OF EMBEDDING AND NOUNS BEING IN A PP OR CPP

V. DISCUSSION

We have shown how data from a reading corpus can be used to investigate reading behavior of human readers in a more or less natural linguistic environment. It might be argued that the text genre represented in the Freiburg Legalese Reading Corpus is not really every-day language. However, the linguistic material included can be seen as an instance of difficult terminological texts that might be encountered in

every-day life. The texts are connected between one another by referential relations, shared topics (or topic changes) and a rhetorical structure. Most of the stimuli used in controlled psycholinguistic experiments do not have these properties. That is why findings from reading corpora might be an attractive way to complement findings from strictly controlled sentence comprehension studies. In the present paper, we showed that the accelerating effect of embedding depth on reading speed can also be shown in the Freiburg Legalese Reading Corpus. Pynte et al. (2008a) suggest that this effect might be an expression of a certain reading strategy. This strategy could lead readers to allocate less attention to portions of text that are less central for the main message of the text. However, we also showed that the acceleration effect varies in magnitude over different parts-of-speech. Nouns seem to be affected the most. Since the hypothesis of Pynte et al. explicitly mentions modifying structures, we divided nouns into two groups: The ones which direct parents in the phrase structure annotation are prepositional phrases and the ones where this is not the case. Nouns, that are directly embedded in a prepositional phrase show an even more pronounced effect of depth of embedding. This is expressed in a significant interaction effect between the depth of embedding and the factor 'in (C)PP'.

To put these findings into perspective, they also have to be discussed in a broader range of psycholinguistic theories. To a certain extent, some findings might be explained by prediction-based theories. Several operationalizations in the field of sentence processing research can be connected to the concept of prediction or predictability. Syntax-based concepts like surprisal (Hale, 2001) or the syntactic constraint score devised by Pynte et al. (2008a, 2009b) have in common that they use syntactic annotations of large-scale corpora to devise metrics of how probable it is that a certain syntactic constituent appears given the preceding syntactic context. Pynte et al. (2008a) derive a semantic constraint score using latent semantic analysis (LSA) to measure the *semantic* distance of any given word to the preceding sentence fragment. Another operationalization of predictability is an empirical one. Kliegl et al. (2004) showed that the probability of a word given its preceding sentence context can be measured via a cloze task. Here, many participants fill in the next word in a sentence given its preceding context. Words, that are very likely to be predicted in such a cloze task, are generally read faster. These are all quite different operationalizations of predictability, but every single one of these concepts is a possible theoretical competitor for the relevance account we based our analyses on. Predictability accounts compete with the relevance account because they predict the same effects of relative position within a sentence and depth of embedding: As a sentence gets longer, the number of possible continuations gets smaller. This has to be thought of as a general effect as it is certainly not true in every single sentence of a language or even a single text.

VI. Outlook

One obvious first step would be to test if the relevance hypothesis accounts for variance in reading times apart from the different predictability accounts introduced very briefly above. However, at the moment we cannot see how the interaction effect of nouns within PPs we showed in the last analysis can be accounted for by predictability approaches. One answer could be that predictability is distributed unevenly over parts-of-speech and phrase types in a way that nouns within PPs are more predictable than all other nouns.

Another important step in providing support for the relevance hypothesis would be to test another hypothesis which follows from the first one. The structures that are embedded deeply and hence get less attention (and are thus read faster) should also be remembered worse. When participants allocate fewer resources to these structures during processing, the mental representations in memory should be less rich or the memory representation decays completely until the reader is finished with the text. Such a hypothesis had to be tested with questionnaires or other appropriate methods after text reading. For the Freiburg Legalese Reading Corpus, such measures are unfortunately not available at the moment.

References

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. http://CRAN.R-project.org/package=lme4 (R package version 1.1-8)

Frank, S. L., Monsalve, I. F., Thompson, R. L. & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior Research Methods, 45,* 1182-1190.

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In P*roceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, S. 159- 166). Pittsburgh, PA: Association for Computational Linguistics.

Heister, J.,Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A. & Kliegl, R.(2011). dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau, 62* (1), 10-20.

Kennedy, A., Hill, R. L. & Pynte, J. (2003). The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movements*. Dundee.

Kliegl, R., Grabner, E., Rolfs, M. & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16* (1/2), 262-284.

Pynte, J., New, B. & Kennedy, A. (2008a). A multiple regression analysis of syntactic and semantic influences in reading normal text. *Journal of Eye Movement Research, 2* (1), 1-11.

Pynte, J., New, B. & Kennedy, A. (2008b). On-line contextual influences during reading normal text: A multiple regression analysis. *Vision Research, 48* (21), 2172-2183.

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria. Zugriff auf http://www.R-project.org/

Tabor, W., Galantucci, B. & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language, 50*, 355-370.

Wolfer, S., Müller-Feldmeth, D., Konieczny, L., Held, U., Maksymski, K., Hansen-Schirra, S., ... Auer, P. (2013). PopSci: A reading corpus of popular science texts with rich multi-level annotations. A case study. In K. Holmqvist, F. Mulvey & R. Johansson (Hrsg.), *Book of Abstracts of the 17th European Conference on Eye Movements*. Lund.